Researchers were interested in the effect that acid has on the growth rate of alfalfa plants. They created three treatment groups in an experiment: low acid, high acid, and control. The alfalfa plants were grown in a Styrofoam cups arranged near a window and the height of the alfalfa plants was measured after five days of growth. The experiment consisted of 5 cups for each of the 3 treatments, for a total of 15 observations.

<table>
<thead>
<tr>
<th></th>
<th>High Acid</th>
<th>Low Acid</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1.30</td>
<td>1.15</td>
<td>0.50</td>
</tr>
<tr>
<td></td>
<td>1.78</td>
<td>1.25</td>
<td>1.27</td>
</tr>
<tr>
<td></td>
<td>2.67</td>
<td>2.25</td>
<td>1.66</td>
</tr>
<tr>
<td></td>
<td>0.30</td>
<td>0.55</td>
<td>1.66</td>
</tr>
<tr>
<td></td>
<td>1.30</td>
<td>0.80</td>
<td>0.80</td>
</tr>
<tr>
<td>$\bar{y}_i$</td>
<td>0.91</td>
<td>1.13</td>
<td>1.77</td>
</tr>
<tr>
<td>n</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

$\mu = 1.269$

We would like to establish if the acid treatments are affecting the alfalfa’s growth. Since we have a numerical response and categorical explanatory variable (> 2 levels) we will use an ANOVA.

What should our hypotheses be?

$H_0$: $\mu_H = \mu_L = \mu_C$

$H_A$: At least one pair of means differ

Last time we mentioned that it is possible to write down a model for each data point

$$y_{ij} = \mu_i + \epsilon_{ij}$$

where $i \in \{H, L, C\}$ is the treatment and $j \in \{1, 2, 3, 4, 5\}$ is the index of the observation within that treatment.

We can rewrite this in terms of the grand mean $\mu$ as follows

$$y_{ij} = \mu + \tau_i + \epsilon_{ij}$$

where $\tau_i = \mu_i - \mu$ is the treatment effect of treatment $i$.

Thinking in terms of the treatment effect we can rewrite our null hypothesis

$$H_0: \mu_H = \mu_L = \mu_C = \mu \quad \Rightarrow \quad H_0: \tau_H = \tau_L = \tau_C = 0$$
One-way ANOVA

Example - Alfalfa

Alfalfa ANOVA Table - Sum Sq

<table>
<thead>
<tr>
<th></th>
<th>df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>1</td>
<td>1.986</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Residuals</td>
<td>12</td>
<td>3.893</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>14</td>
<td>5.879</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[
SST = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \mu)^2 \\
= (1.3 - 1.269)^2 + (1.15 - 1.269)^2 + \cdots + (0.80 - 1.269)^2 = 5.879
\]

\[
SSG = \sum_{i=1}^{k} n_i (\mu_i - \mu)^2 \\
= 5 \times (0.91 - 1.269)^2 + 5 \times (1.13 - 1.269)^2 + 5 \times (1.768 - 1.269)^2 = 1.986
\]

\[
SSE = SST - SSG = 3.893
\]

Alfalfa ANOVA Table - DF

<table>
<thead>
<tr>
<th></th>
<th>df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>2</td>
<td>1.986</td>
<td>0.993</td>
<td>3.061</td>
<td>0.0843</td>
</tr>
<tr>
<td>Residuals</td>
<td>12</td>
<td>3.893</td>
<td>0.324</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>14</td>
<td>5.879</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[
df_T = n - 1 = 15 - 1 = 14 \\
df_G = k - 1 = 3 - 1 = 2 \\
df_E = n - k = 15 - 3 = 12
\]

Based on these results we fail to reject \( H_0 \), and there is not sufficient evidence to suggest that at least one pair of mean growth values are significantly different.

Randomized Block Design

Blocking

Random Sampling / Assignment

Random sampling removes nuisance factors/variables (things that affect your outcome that you are not interested in).

Imagine we are interested in exploring whether increasing the dosage of a Statin will reduce the risk of a heart attack. We randomly sample patients already on a Statin and randomly assign them to either maintain their current dosage or increase their dosage by 20%.

- Possible that some of the patients in this sample may have had a previous heart attack,
- Significant risk factor for a future heart attack
- Their presence may alter our outcome
- Control for this effect by excluding them

Ideally random sampling / assignment ensure that in the long run these nuisance factors show up with equal frequency in all treatment levels and as such their effect(s) will cancel out.
Randomized Block Design

Blocking

Why do we bother with controls then? Because they help reduce noise/uncertainty in the data.

Types of Controls
- Exclusion
  - Works if the number of patients with a previous heart attack is low
  - Can only exclude so many nuisance factors
  - Restricts generalizability
- Blocking
  - Samples grouped into homogeneous blocks where the nuisance factor(s) are held constant
  - Variation within the block should be less than the variation between blocks
  - Previous heart attack block and a no previous heart attack block
  - Randomized treatment assignment within each block

“Block what you can; randomize what you cannot.”

Types of Controls
- Exclusion
  - Works if the number of patients with a previous heart attack is low
  - Can only exclude so many nuisance factors
  - Restricts generalizability
- Blocking
  - Samples grouped into homogeneous blocks where the nuisance factor(s) are held constant
  - Variation within the block should be less than the variation between blocks
  - Previous heart attack block and a no previous heart attack block
  - Randomized treatment assignment within each block

“Block what you can; randomize what you cannot.”

Blocked Alfalfa

We will only consider the simplest case of randomized block design where each block contains only one observation of each treatment.

<table>
<thead>
<tr>
<th></th>
<th>High Acid</th>
<th>Low Acid</th>
<th>Control</th>
<th>Block Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Block 1</td>
<td>1.30</td>
<td>1.78</td>
<td>2.67</td>
<td>1.917</td>
</tr>
<tr>
<td>Block 2</td>
<td>1.15</td>
<td>1.25</td>
<td>2.25</td>
<td>1.550</td>
</tr>
<tr>
<td>Block 3</td>
<td>0.50</td>
<td>1.27</td>
<td>1.46</td>
<td>1.077</td>
</tr>
<tr>
<td>Block 4</td>
<td>0.30</td>
<td>0.55</td>
<td>1.66</td>
<td>0.837</td>
</tr>
<tr>
<td>Block 5</td>
<td>1.30</td>
<td>0.80</td>
<td>0.80</td>
<td>0.967</td>
</tr>
<tr>
<td>Trmt mean</td>
<td>0.910</td>
<td>1.130</td>
<td>1.768</td>
<td>1.269</td>
</tr>
</tbody>
</table>

\[ \mu = 1.269 \]

Block Data Model

When employing blocks we can think of each data point as

\[ y_{ijk} = \mu + \tau_i + \beta_j + \epsilon_{ijk} \]

where

- \(\tau_i\) is the treatment effect for treatment \(i\)
- \(\beta_j\) is the block effect of block \(j\)
- \(\epsilon_{ijk}\) is the residual of observation \(k\) in block \(j\) with treatment \(i\)

this is very similar to the one-way anova model we saw previously with the addition of the \(\beta_j\)s.
Randomized Block Design

Blocked Alfalfa

Randomized Block ANOVA Table

With the introduction of the blocks there are now two hypotheses we would like to evaluate:

\[ H_0(\text{treatment}) : \tau_T = \tau_L = \tau_C = 0 \]
\[ H_0(\text{block}) : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0 \]

In order to test these hypotheses we will build on the ANOVA table we have been using.

\[
\begin{array}{llllll}
\text{df} & \text{Sum Sq} & \text{Mean Sq} & F \text{ value} & \text{Pr}(>F) \\
\hline
\text{Group} & df_G & SS_G & MSG & F_G & \\
\text{Block} & df_B & SS_B & MS_B & F_B & \\
\text{Error} & df_E & SS_E & MSE & \\
\text{Total} & df_T & SST & & & \\
\end{array}
\]

Randomized Block ANOVA Table - Alfalfa

We already know some of the values from our previous one-way ANOVA, and it is easy to find the other \textit{df} values.

\[
\begin{array}{llll}
\text{df} & \text{Sum Sq} & \text{Mean Sq} & F \text{ value} \\
\hline
\text{Group} & 2 & 1.986 & 0.993 & MSG/MSE \\
\text{Block} & 4 & \sum_{j=1}^{b} m_j (\mu_{i,j} - \mu)^2 & SSB/df_B & MSB/MSE \\
\text{Error} & 8 & SST - SS_G - SS_B & SSE/df_E & \\
\text{Total} & 14 & 5.879 & & \\
\end{array}
\]

Sum of Squares Blocks

\[ SS_B = \sum_{j=1}^{b} m_j (\mu_{i,j} - \mu)^2 \]

\[
\begin{array}{cccccc}
\text{Block} & \text{High Acid} & \text{Low Acid} & \text{Control} & \text{Block Mean} \\
\hline
\text{Block 1} & 1.30 & 1.78 & 2.67 & 1.917 \\
\text{Block 2} & 1.15 & 1.25 & 2.25 & 1.550 \\
\text{Block 3} & 0.50 & 1.27 & 1.46 & 1.077 \\
\text{Block 4} & 0.30 & 0.55 & 1.66 & 0.837 \\
\text{Block 5} & 1.30 & 0.80 & 0.80 & 0.967 \\
\text{Trmt mean} & 0.910 & 1.130 & 1.768 & \\
\hline
\end{array}
\]

\[ SS_B = 3 \times (1.917 - 1.269)^2 + 3 \times (1.550 - 1.269)^2 \]
\[ + 3 \times (1.077 - 1.269)^2 + 3 \times (0.837 - 1.269)^2 \]
\[ + 3 \times (0.967 - 1.269)^2 \]
\[ = 1.260 + 0.237 + 0.111 + 0.560 + 0.274 = 2.441 \]
Completing the table

<table>
<thead>
<tr>
<th></th>
<th>df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group</td>
<td>2</td>
<td>1.986</td>
<td>0.993</td>
<td>5.471</td>
</tr>
<tr>
<td>Block</td>
<td>4</td>
<td>2.441</td>
<td>0.6103</td>
<td>3.362</td>
</tr>
<tr>
<td>Error</td>
<td>8</td>
<td>1.452</td>
<td>0.1815</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>14</td>
<td>5.879</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Calculating P-values

The two F values that we have calculated can be used to evaluate the two hypotheses we started with.

- **Treatment effect**
  
  \[ H_0 : \tau_H = \tau_L = \tau_G, \quad H_A : \text{At least one pair of treatment effects differ} \]

- **Block effect**
  
  \[ H_0 : \beta_1 = \beta_2 = \ldots = \beta_5, \quad H_A : \text{At least one pair of block effects differ} \]

To calculate the P-value for each hypothesis we use \( F_G \) and \( F_B \) respectively to find \( P(> F) \) for an F distribution with the appropriate degrees of freedom.

**Treatment Effect**

We have calculated that \( F_G = 5.471 \), to find the P-value we need to the probability of observing a value equal to or larger than this from an F distribution with 2 and 8 degrees of freedom.

Using R we find that

```r
pf(5.471, df1=2, df2=8, lower.tail=FALSE)
## [1] 0.03181681
```

Therefore, \( P(> F_G) = 0.0318 \), which leads us to reject \( H_0 \) - there is sufficient evidence to suggest that at least one pair of treatment effects differ.

**Block Effect**

Similarly, we have \( F_B = 3.362 \) and to find the P-value we need to the probability of observing a value equal to or larger than this from an F distribution with 4 and 8 degrees of freedom.

Using R we find that

```r
pf(3.362, df1=4, df2=8, lower.tail=FALSE)
## [1] 0.06790077
```

Therefore, \( P(> F_B) = 0.0679 \), which leads us to fail to reject \( H_0 \) - there is not sufficient evidence to suggest that at least one pair of block effects differ.
How did blocking change our result?

- One-way ANOVA

<table>
<thead>
<tr>
<th></th>
<th>df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>2</td>
<td>1.986</td>
<td>0.993</td>
<td>3.061</td>
<td>0.0843</td>
</tr>
<tr>
<td>Residuals</td>
<td>12</td>
<td>3.893</td>
<td>0.324</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>14</td>
<td>5.879</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Randomized Block ANOVA

<table>
<thead>
<tr>
<th></th>
<th>df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>P(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group</td>
<td>2</td>
<td>1.986</td>
<td>0.993</td>
<td>5.471</td>
<td>0.0318</td>
</tr>
<tr>
<td>Block</td>
<td>4</td>
<td>2.441</td>
<td>0.6103</td>
<td>3.362</td>
<td>0.0679</td>
</tr>
<tr>
<td>Error</td>
<td>8</td>
<td>1.452</td>
<td>0.1815</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>14</td>
<td>5.879</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Blocking decreases $df_E$, which increases $MSE$ (bad). Blocking also decreases $SSE$, which decreases $MSE$ (good).

From Randomized Block to Two-way ANOVA

All of the approaches we have just learned to handle blocking will also apply in the case where we would like to assess the effect if a second factor on our outcome variable.

Instead of examining treatment and block effects we instead examine two treatment effects. None of the procedures or calculations change, only what we call things.

Two-way ANOVA Model

When employing two-way ANOVA we can think of each data point as

$$y_{ijk} = \mu + \tau_i + \beta_j + \epsilon_{ijk}$$

where

- $\tau_i$ is the effect of level $i$ of treatment 1
- $\beta_j$ is the effect of level $j$ of treatment 2
- $\epsilon_{ijk}$ is the residual of observation $k$ in with treatment 1 level $i$ and treatment 2 level $j$

this is exactly the same as the randomized block ANOVA model except the $\beta_j$s now refer to the effect of the second factor.

Example - Spruce Moths

A scientist is interested in efficacy of various lure types in attracting Spruce moths to a trap. They are also interested in the effect of location of the trap on its efficacy as well.

Data to the right reflects the number of moths caught.

Factor 1 is the lure type (3 levels)
Factor 2 is the location (4 levels)
There are 5 observations per condition

From Understandable Statistics, 7e
### Example - Moths

**Mean caught by Treatment**

<table>
<thead>
<tr>
<th></th>
<th>Ground</th>
<th>Lower</th>
<th>Middle</th>
<th>Top</th>
<th>Lure Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chemical</td>
<td>19.20</td>
<td>36.00</td>
<td>31.80</td>
<td>23.00</td>
<td>27.50</td>
</tr>
<tr>
<td>Scent</td>
<td>17.00</td>
<td>32.80</td>
<td>27.80</td>
<td>21.40</td>
<td>24.75</td>
</tr>
<tr>
<td>Sugar</td>
<td>21.00</td>
<td>31.20</td>
<td>33.40</td>
<td>25.60</td>
<td>27.80</td>
</tr>
<tr>
<td>Loc Mean</td>
<td>19.07</td>
<td>33.33</td>
<td>31.00</td>
<td>23.33</td>
<td>26.68</td>
</tr>
</tbody>
</table>

---

**Example - Spruce Moths - Hypotheses**

Similar to the randomized block ANOVA, we have two hypotheses to evaluate (one for each factor).

**Lure effect:**

\[ H_0 : \tau_{Ch} = \tau_{Sc} = \tau_{Su} \quad H_A : \text{at least one pair of } \tau \text{s differ} \]

**Location effect:**

\[ H_0 : \beta_G = \beta_L = \beta_M = \beta_T \quad H_A : \text{at least one pair of } \beta \text{s differ} \]

---

**Example - Spruce Moths - ANOVA Table**

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lure</td>
<td>0</td>
<td>0.3859</td>
<td></td>
<td>0.3859</td>
<td></td>
</tr>
<tr>
<td>Location</td>
<td>1981.38</td>
<td>0.0000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Residuals</td>
<td>5242.98</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Conclusions:**

- Fail to reject \(H_0\) (Lure), there is not sufficient evidence to suggest the different lures have an effect.
- Reject \(H_0\) (Location), there is sufficient evidence to suggest the locations have an effect.
Difference between a blocking variable and a factor

We have just seen that computationally the two are treated the same when conducting an ANOVA.

What then is the difference?

- Factors are conditions we impose on the experimental units.
- Blocking variables are characteristics that the experimental units come with.

Example - Lighting

A study is designed to test the effect of type of light on exam performance of students. 180 students are randomly assigned to three classrooms: one that is dimly lit, another with yellow lighting, and a third with white fluorescent lighting and given the same exam.

What are the factor(s) and/or block(s) for this experiment? What type of ANOVA would be appropriate?

The researcher also believes that light levels might have a different effect on males and females, so wants to make sure both genders are represented equally under the different light conditions.

After this modifications what are the factor(s) and/or block(s) for this experiment? What type of ANOVA would be appropriate?